로짓(Logit) 모델을 이용한 날씨요소와 송전선로 고장의 다중회귀분석

신동석* · 이윤호* · 김진오* · 이백석** · 방민재**
*한양대학교 전기공학과 · **한국전력공사

Multiple Regression Analysis between Weather Factor and Line Outage using Logit Model

Dong-Suk Shin* · Youn-Ho Lee* · Jin-O Kim* · Baek-Seok Lee** · Min-Jae Bang**
*Dept. of EE, Hanyang University · **KEPCO

Abstract - This Paper investigates the effect of weather factors(such as winds, rain, snows, temperature, clouds and humidity) on transmission line outages. The result shows that weather variables have significant effects on the transmission line historical outages and the relationship between them is nonlinear. Multiple regression analysis using Logit model is proved to be appropriate in forecasting line failure rate in KEPCO systems. It could also provide system operators with useful informations about system operation and planing.

1. 서 론

오늘날 산업 및 사회의 성장으로 인한 전력수요의 급격한 증가는 전력시스템을 복잡하고 거대한 시스템으로 만들었다. 그로인해 안정적인 전력공급과 발전비용의 감소 및 전력공급 신뢰도의 향상과 같은 여러 문제가 대두되었다. 그러나 이렇게 거대해진 시스템을 정확하게 분석한다는 것은 대단히 어려운 일이다. 따라서 시스템의 분석 및 예측에 통계적 해석방법을 사용하려 한다.

지금까지의 전력산업은 나라에서 공익을 목적으로 설계 및 운영되어왔지만, 현재는 경쟁체제로 변화하고 있다. 따라서 전력 운영자들은 경제적인 이윤을 창출하기 위해 시스템을 더욱 효율적으로 운영해야만 한다. 여기에 많은 영향을 미치는 것이 상정사고인데 이것은 시스템 운영과 계획에 중요한 기준이 되기 때문이다. 이런 상정사고에 영향을 주는 확률적인 요인들 중에서 본 논문에서는 날씨요인에 대해 평가할 것이다. 왜냐면 전력설비들은 대부분이 외부에 많이 설치되어 있기 때문에 날씨의 영향을 많이 받기 때문이다. 그러나 지금까지는 이런 날씨의 영향을 시스템에 잘 반영하지 못하고 있다. 따라서 날씨가 시스템에 미치는 영향에 대하여 평가하고 이를 반영한다면 더 효율적인 운영이 가능할 것이다. 그리고 전력시스템에서 송전선로는 대표적인 외부설비로 넓은 지역에 걸쳐 설치되기 때문에 그 영향이 가장 많을 것으로 판단된다. 그러므로 이런 송전선로에 미치는 날씨의 영향에 대해 평가하려고 한다.

본 논문에서는 1994년에서 2003년까지의 9개 관리처 데이터와 기상청 데이터를 사용하여 기상요소들에 따른 고장률을 계산하고, 이것을 이용해서 로짓(Logit) 모델로 날씨요소들을 설명변수로 하는 다중회귀모델을 제안하려고 한다.

2. 본 론

2.1 날씨에 따른 송전선로 고장률

날씨와 송전선로 고장과의 관계를 알아보기 위해서는 고장이 발생한 날짜의 풍속, 강수량, 적설량, 기온 및 상대습도와 같은 기상데이터와 고장데이터가 필요하다. 이런 방대한 데이터 처리를 위해 데이터베이스를 구축하였다. 여기서 나온 데이터를 이용하여 고장률을 구할 수 있은데, 그 표현은 식 (1)과 같이 나타낼 수 있다.[2,4]

$$\text{고장률} = \frac{\text{고장발생횟수}}{\text{총 시간}} \tag{1}$$

여기서, 총 시간 : 날씨요소 값에 따른 총 시간[hour]
고장발생 횟수: 총 시간 동안의 고장횟수

2.2 날씨요소별 송전선로 고장률

1994년에서 2003년까지의 9개 관리처 데이터와 기상청 데이터를 사용하여 사고발생건수와 기상요소별 크기에 따른 총 시간데이터를 식 (1)에 입력하여 고장률을 계산하면 그림 1과 같다.[1]

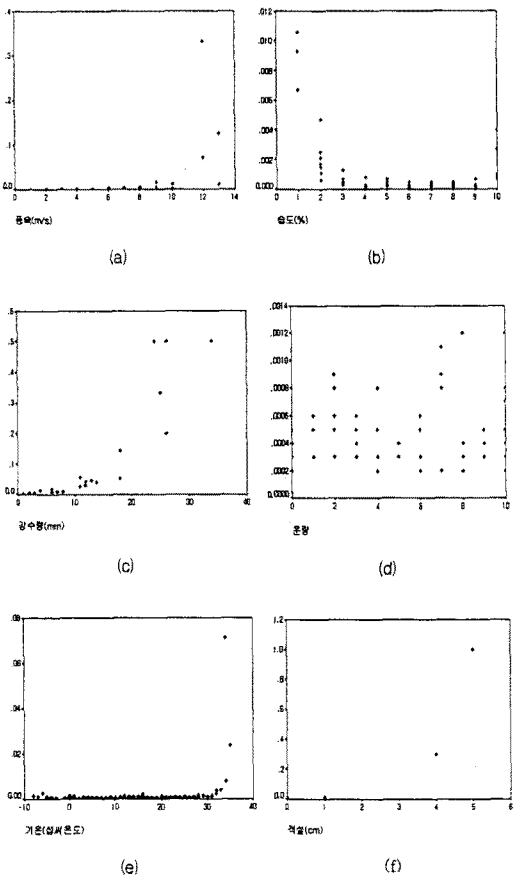

그림 1. 각 날씨요소들과 고장확률

그림 1과 같이 날씨요소와 고장률의 관계는 비선형이다. 그리고 날씨요소 값이 적은 평상날씨에서는 고장률이 거의 직선으로 매우 작지만, 반대로 가혹날씨로 변하면 고장률이 급격히 증가하는 것을 볼 수 있다.

2.3 로짓모델

로짓모형은 종속변수가 0과 1인 값만을 가질 수 있는 가변수인 경우에 가장 널리 적용할 수 있는 분석모형이다. 전력계통의 상정사고도 이와 같이 정상상태를 0과 사고 발생시를 1로 놓고 볼 수 있기 때문에 이 모형의 적용이 가능하다. 이것은 다음 식(2)의 관계로 설명 된다.

$$P_i = E[Y=1 \mid X_i] = \alpha + \beta X_i \tag{2}$$

여기서 Y=1이 되는 확률은 P이고, Y=0이 되는 확률은 1-P가 된다. 그리고 이 모형에 대해서 보편적으로 받아지는 함수는 누적분포함수(Cumulative Distribution Function : CDF)이며, 이것을 근거한 분석모형이 로직모형이다.

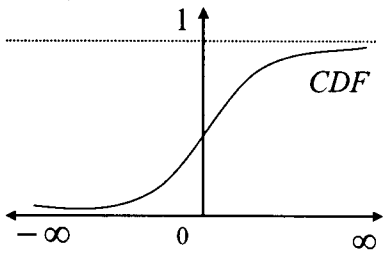

그림 2. 누적분포함수(CDF)

따라서 식(2)은 다음의 식(3)으로 표현된다.

$$P_i = E[Y=1 \mid X_i] = \frac{e^{\alpha+\beta X_i}}{1+e^{\alpha+\beta X_i}} = \frac{1}{1+e^{-Z_i}} \tag{3}$$

단, $Z_i = \alpha + \beta X_i$

식 (3)에서 $Z_i \rightarrow +\infty$이면 $P_i \rightarrow 0$이고, $Z_i \rightarrow -\infty$이면 $P_i \rightarrow 1$이 되어 예측되는 범의는 0과 1사이가 된다. 그러나 P_i는 비선형관계이기 때문에 식 (3)을 선형관계로 변환한 후 최소자승법을 이용하여 추정해야 한다. 다음은 그 과정이다.

$$1 - P_i = \frac{1}{1+e^{Z_i}} \tag{4}$$

$$\frac{P_i}{1-P_i} = \frac{1+e^{Z_i}}{1+e^{-Z_i}} = e^{Z_i} \tag{5}$$

식 (5)의 양변에 log를 취하면 식 (6)과 같은 선형관계식을 얻을 수 있으며, 이것을 로직 모형이라고 한다.

$$L_i = ln\left[\frac{P_i}{1-P_i}\right] = Z_i = \alpha + \beta X_i \tag{6}$$

2.3.1 로짓모델을 이용한 다중회귀분석

송전선로 고장의 다중회귀분석을 위한 설명변수인 날씨요소들에서 상관관계가 적은 운량과 데이터가 적은 적설량은 빠진다. 그리고 설명변수 중에서 상관관계가 가장 높은 강수량을 기준으로 P(고장률)를 구하고, 나머지 요소들은 그때의 평균값들을 사용하려 한다. 표 1은 그 예이며, 지면관계상 중간부분은 생략하였다.

표 1. 강수량에 따른 Sample Data

강수량	P	logit	풍속	기온	습도
0	0.0003	-8.1114	3.5122	15.8841	5.2317
0	0.0003	-8.1114	3.5122	15.8841	5.2317
0	0.0003	-8.1114	3.5122	15.8841	5.2317
0	0.0003	-8.1114	3.5122	15.8841	5.2317
0	0.0003	-8.1114	3.5122	15.8841	5.2317
0	0.0001	-9.2102	3.5122	15.8841	5.2317
0	0.0002	-8.517	3.5122	15.8841	5.2317
0	0.0002	-8.517	3.5122	15.8841	5.2317
1	0.0021	-6.1637	3.5385	19.6154	8.2308
1	0.0007	-7.2637	3.5385	19.6154	8.2308
⋮	⋮	⋮	⋮	⋮	⋮
26	0.2	-0.3863	2	22.5	8.5
26	0.5	0	2	22.5	8.5
34	0.5	0	2	18	9

단, P는 강수량에 따른 고장률이 되고, 식 (7)과 같다.

$$P_i = \frac{n_i}{N_i} \tag{7}$$

N_i : 날씨요소의 크기에 따른 총 시간

n_i : 날씨요소의 크기에 따른 송전선로 고장 횟수

여기서 P_i의 분포가 정규분포를 따르지 않기 때문에 회귀분석 값으로 적정하지 않다. 따라서 로짓모델을 이용하여 정규분포를 따르는 logit 변수인 $\ln\left(\frac{P_i}{1-P_i}\right)$ 값으로 바꾸었다.

또한, 식 (6)을 다중회귀로 확장하고, 그림 1에서 보는 바와 같이 비선형관계이기 때문에 각 날씨요소들의 제곱을 설명변수로 추가하면, 식 (9)와 같이 표현된다.

$$\ln\left(\frac{P_i}{1-P_i}\right) = \alpha_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \beta_5 X_3 + \beta_6 X_3^2 + \beta_7 X_4 + \beta_8 X_4^2 \tag{9}$$

(X_1 : 강수량 X_2 : 풍속 X_3 : 기온 X_4 : 습도)

2.4 검정 통계치

회귀모형이 추정되면 자료들을 얼마나 잘 설명하는가를 검토해야 한다. 따라서 본 논문에서는 다음과 같은 척도들을 사용하여 검토한다.

(1) 결정계수(Coefficient of determination : R^2)

회귀모형이 표본자료와 얼마나 적합한지를 나타내는 값으로 1에 가까울수록 적합하다.

(2) DW(Durbin-Watson)

최소자승법을 사용하여 회귀계수를 추정할 때 잔차는 서로 독립적이라는 가정을 하고 있다. 따라서 추정한 계수가 적정하다고 말하기 위해서는 잔차항들에서 자기상관관계(auto-correlation)가 존재하면 안 된다. 이 통계량은 이것을 검증하는데 자기상관이 양으로 증가할수록 감

소하고, 반대로 음으로 증가할수록 증가하게 된다. 그리고 자기상관이 없는 경우 2에 가까운 값이 나온다.

(3) SE(Standard Error)
표준오차(SE)는 데이터가 얼마나 회귀모델에 집중되어 있는가를 나타내는 지표이다.

(4) t-values and t-probability
이 값들은 계수(coefficient) 값이 0인지 아닌지를 나타내어주는 통계량이다.

2.5 다중회귀분석 결과

PcGive(통계프로그램)을 이용하여 계산해 보면 식 (10)과 같은 결과를 얻을 수 있다.

$$\ln\left(\frac{P_i}{1-P_i}\right) = -17.194 + 0.45548X_1 - 0.008031X_1^2 - 0.80749X_2 + 0.069463X_2^2 + 0.11107X_3 - 0.0067007X_3^2 + 2.989X_4 - 0.17906X_4^2 \quad (10)$$

표 2. 모델의 적정도

R^2	DW
0.96856	2.08

R^2 값에서 보듯 로짓모델을 이용해서 변화를 96.85% 설명할 수 있다. 또한, DW도 2에 가까우므로 자기상관관계가 없다고 할 수 있다. 각 변수에 대한 계수와 검정통계치를 정리하면 다음과 같다.

표 3. 로짓모델의 결과

variable	coefficient	std.Error	t-value	t-prob
constant	-17.194	3.4771	-4.945	0.0000
강수량	0.45548	0.007811	5.854	0.0000
강수량^2	-0.0080310	0.0023494	-3.418	0.0017
풍속	-0.80749	0.28312	-2.852	0.0075
풍속^2	0.069463	0.028774	2.414	0.0217
기온	0.11107	0.093250	1.191	0.2424
기온^2	-0.0067007	0.0034296	-1.954	0.0595
습도	2.9890	1.0415	2.870	0.0072
습도^2	-0.17906	0.073661	-2.431	0.0208

표 3에서 추정치의 t값들이 대부분 2가 넘으므로 95%의 신뢰수준에서 유의성이 있다. 특히, 제곱한 변수의 t값도 유의성이 있으므로 비선형임을 추정할 수 있다. 그런데 식 (10)은 고장률 $\hat{P}_i$에 대한 식이 아니므로 식 (3)을 이용하여 바꾸면, 식 (11)과 같은 송전선로 고장에 의한 추정치를 얻을 수 있었다.

$$\hat{P}_i = \frac{1}{1+e^{-\hat{Z}_i}} \quad (11)$$

여기서,

$$\hat{Z}_i = -17.194 + 0.45548X_1 - 0.008031X_1^2 - 0.80749X_2 + 0.069463X_2^2 + 0.11107X_3 - 0.0067007X_3^2 + 2.989X_4 - 0.17906X_4^2$$

(X_1 : 강수량 X_2 : 풍속 X_3 : 기온 X_4 : 습도)

로짓모델을 이용하여 회귀분석 결과로 나온 $\hat{P}_i$(추정치)와 P_i(실제 고장률)를 비교해 보면 그림 3과 같다.

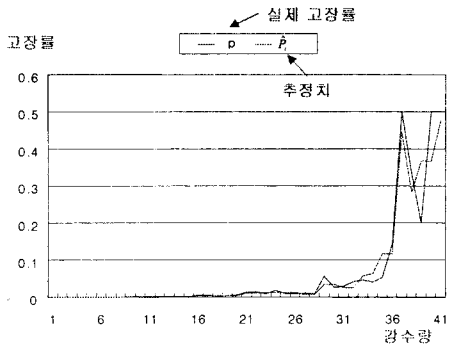

그림 3. 실제치와 추정치 비교

따라서 X_1(강수량), X_2(풍속), X_3(기온) 및 X_4(습도)를 예상할 수 있다면, 식 (11)을 이용하여 $\hat{Z}_i$을 계산하여 고장률 추정치를 구할 수 있다.

3. 결 론

본 논문은 날씨가 송전선로 사고에 얼마나 영향을 미치는가를 통계적인 방법으로 분석함으로써 시스템의 운영 및 계획에 도움을 주고자 하였다. 사용된 고장데이터는 1994년에서 2003년까지의 10년간의 KEPCO의 송전선로 고장데이터와 그 기간의 기상청 데이터이다. 이것을 이용하여 각 날씨요소별 고장률이 얼마이고, 어떤 경향을 가지는지 알아보았다. 그리고 다중회귀분석을 하기 위해 로짓(Logit) 모델을 사용하여 상관관계가 높은 강수량을 기준으로 풍속, 기온 및 습도를 설명변수로 고장률을 예측할 수 있는 다중회귀모델을 제안하였다.

따라서 이 모델을 이용하여 기상의 변화에 대하여 고장을 미리 예측할 수 있을 것이고, 이를 전력계통운영에 반영할 수 있다면 더욱 경제적인 운영이 가능할 것으로 사료된다.

감사의 글

본 연구는 전력연구원의 연구지원(기금-119J03PJ03)에 의해 수행되었음.

[참 고 문 헌]

[1] J. McDaniel, C. Williams and A. Vestal, "Lightning and Distribution Reliability-A Comparison of Three Utilities", *IEEE*, 2003.
[2] C.W. Williams and Jr. PE CPQ, "Weather Normalization of Power System Reliability Indices", *IEEE*, 2003.
[3] 김상익, 서한손, 안병진, 여성철, 이석구, "통계학의 이해와 응용", 민영사, 1999.
[4] 배현웅, "통계학의 기초와 활용기법", 교우사, 2002
[5] M.J. Crowder, A.C. Kimber, R.L. Smith and T.J. Sweeting, "Statistical Analysis of Reliability Data", *Chapman & Hall*, 1991.
[6] Dale J. Poirier, "A Bayesian analysis of nested logit models", *Journal of Econometrics*, 1996.
[7] Beals, Ralph E., "Statistics for Economists", *Rand McNally*, 1972.
[8] Pindyck, Robert S., Daniel L. Rubinfeld, "Econometric Models and Economic Forecasts", *McGraw-Hill*, 1991.
[9] Huang, David S., "Regression and Econometric Methods", *John Wiley & Sons*, 1964.
[10] Gujarati, Damodar N, "Basic econometrics", *McGraw-Hill*, 2003.